AN ABSTRACT OF THE THESIS OF

Rienk S. Venema for the degree of Master of Science in
Mathematics presented on May 26, 1994.

Title: How the Perceptron Reacts on Non-Separable Classification Problems

Abstract approved: _____

Robert M. Burton

Neural networks are models which have been developed to simulate the anatomy of the nervous system. The connection between the elements of these networks, the so called artificial neurons, is similar to the connection between the biological neurons. In developing neural networks people are trying to create systems which have the same computational and communication properties as the brain.

On the basis of the things we know from neurophysiology the first models for the neural networks are developed. One of these networks was the perceptron, which is one of the most used neural networks. In this thesis we'll study this special neural network. When the input vectors of the perceptron can be linearly separated into two catagories, this network can be trained to correctly classify these input vectors. However in most practical cases the linearly separability assumption isn't satisfied. That's why the main part of this study is devoted to the case where the input vectors aren't linearly separable.

HOW THE PERCEPTRON REACTS ON NON-SEPARABLE CLASSIFICATION PROBLEMS

by

Rienk S. Venema

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Master of Science

Completed May 26, 1994

Commencement June 1995

APPROVED:

Redacted for Privacy

Professor of Mathematics in charge of major

Redacted for Privacy

Chairman of department of Mathematics

Redacted for Privacy

Dean of Graduate School

Date thesis is presented May 26, 1994

Typed by researcher for Rienk S. Venema

Table of Contents

List of Figures

Chapter 1

Introduction

It has been long a desire of some people to create machines that think and reason like people. One of the approaches that has been developed the last forty years is the theory of neural networks. In developing this theory, people are trying to model the structure of the brain by systems which have the same computational and communication properties as the brain.

A main part of the research has been devoted to study how the braincells, the so called neurons, work together because the neurons are responsible for the communication and computational properties of the brain. Unfortunately this information is not complete; a big part of the action of the brain is still a mystery. But every year new data becomes available and the models are adjusted.

In chapter 2 we'll discuss the action of the biological neuron. On the basis of the things we know from neurophysiology, the first models for the artificial neuron were developed, which we'll describe in chapter 3. The artificial neurons are the elements of the neural networks, that basically have the same properties as the biological neurons. The connection between the artificial neurons is similar to the anatomy of the nervous system. In chapter 4 we'll study one of the most used neural network models, namely the perceptron, which has been developed in the fifties.

A fundamental theorem, which will be proved in chapter 4, is the perceptron convergence theorem. This theorem can only be proved under the assumption that

the input vectors of the network can be linearly separated into two catagories. But this is a strong assumption and in most practical cases this assumption isn't satisfied. That's why we'll study in chapter 5 how the network reacts on input vectors which aren't linearly separable. We'll study in particular in chapter 6 the case where the input vectors are normally distributed.

Chapter 2

The biological neuron

2.1 Introduction

A human brain contains over one hundred billion computing elements, called neurons. The neurons communicate throughout the body by way of nerve fibers that make perhaps one hundred trillion connections called synapses. This network of neurons is responsible for all the phenomena that we call thought, emotion and cognition. The exact manner in which this is accomplished is little understood, but much of the physiological structure has been mapped and certain functional areas are gradually yielding to determined research.

The neuron is the fundamental building block of the nervous system. It is a cell similar to all cells in the body; however, certain critical specializations allow it to perform all of the computational and communication functions within the brain.

As shown in figure 1 the neuron consists of three sections:

- the cell body(with the cellular membrane)
- the dendrites
- the axon

Functionally, the dendrites receive the signals from other cells at connection points called synapses.

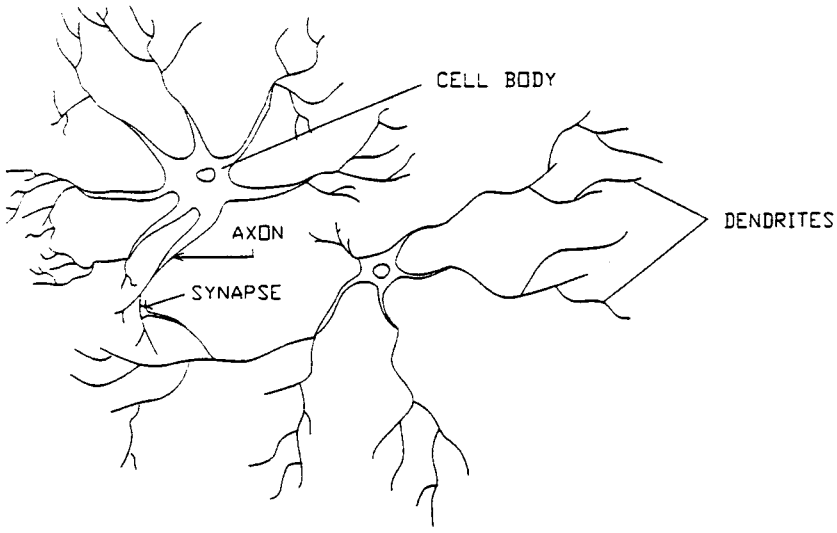

Figure 1. Components of a neuron

2.2 The cell body

The neurons in the adult brain do not regenerate: they must last a lifetime. This means that all of the components must be continuously replaced and the materials renewed as needed. Most of these maintenance activities take place in the cell body, where a wide variety of complex molecules are produced.

The outer membrane of the neuron's cell body has the unique capability of generating nerve impulses, a vital function of the nervous system and central to its computational abilities.

2.3 The dendrites

Most input signals from other neurons enter the cell by way of the dendrites. On the dendrites are synaptic connections where signals are received, usually from other axons. Usually there is no physical or electrical connection made at the synapse. Instead a narrow gap called the synaptic cleft seperates the dendrite from the transmitting axon. See figure 2. Specialized chemicals that are released by the axon into the synaptic cleft diffuse across to the dendrite. These chemicals, called neurotransmitters, enter the cell body. Some of the neurotransmitters are excitatory and tend

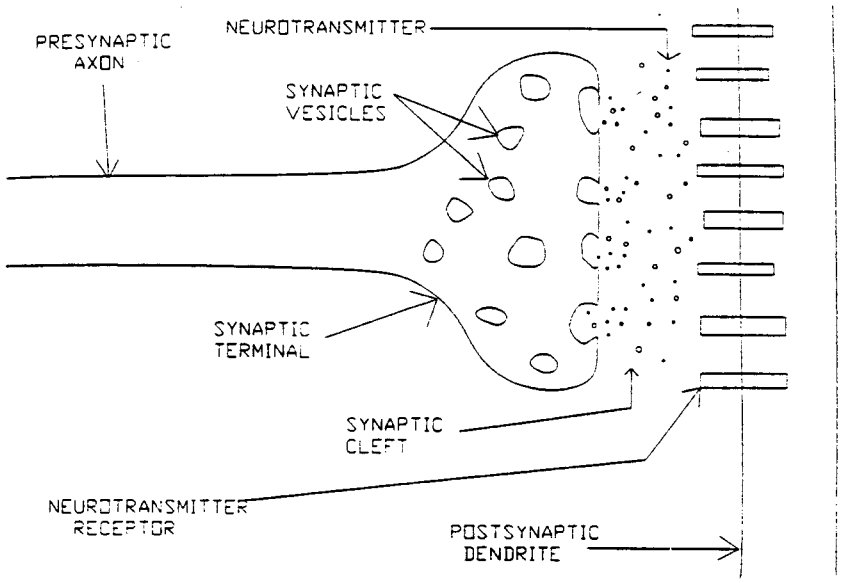

Figure 2. The synapse

to cause the cell to "fire" and produce an output pulse. Others are inhibitory and tend to suppress such a pulse. The neurotransmitters are combined to one signal. The resulting action of the synapse will be called either excitatory or inhibitory.

The cell body combines the signals received over its dendrites and, if their resultant signal is above its threshold, a pulse is produced that propagates down the axon to other neurons(this is called the firing of a neuron).

2.4 The axon

An axon may be as short as 0.1 millimeter or it can exceed 1 meter in length. Near its end, the axon has multiple branches, each terminating in a synapse, where the signal is transmitted to another neuron through a dendrite. In this way a single neuron can generate a pulse that will activate or inhibit hundreds or thousands of other neurons. It is this high degree of connectivity that gives the neuron its computational power.

The axon contains at the end of its branches spherical structures called synaptic vesicles, each of which contains a great number of neurotransmitters. When a nerve impulse arrives down the axon, some of these vesicles release their contents into the synaptic cleft, thereby initiating the process of interneuron communication.

2.5 The cellular membrane

Communication in the brain is of two types: chemical signals across the synapses and electrical signals within the neuron. It is the complex action of the membrane that creates the cell's ability to produce and transmit both kind of signals.

The cell's internal sodium concentration is ten times lower than its surroundings and its potassium concentration is ten times higher. These concentrations tend to equalize through leaks in the cell due to pores in the membrane. To maintain the necessary concentrations, a membrane protein, called a sodium pump, continuously passes sodium out of the cell and potassium into the cell.

A neuron may have millions of such pumps, moving hundreds of millions of ions in and out of the cell each second. The potassium concentration within the cell is further increased by the presence of a large number of permanently open potassium channels; that is, there are proteins that pass potassium ions readily into the cell, but inhibit the passage of sodium. The combination of these two mechanisms is responsible for creating and maintaining the dynamical chemical equilibrium in the neuron.

The ionic concentration gradient across the cell membrane causes the interior of the cell to assume an electrical potential of -70 millivolts relative to its surroundings. When the synaptic inputs have reduced this level to approximately -50 millivolts then the sodium and potassium flows are suddenly reversed. At this point, sodium channels are opened, allowing sodium to flood into the cell. Shortly after opening, the sodium channels close and potassium channels open. This allows potassium to flow out of the cell.

Within a millisecond the interior of the cell becomes +50 millivolts. This results in the firing of the cell and it sends an output impulse through the axon. Finally the neurotransmitters are released into the synaptic cleft and the process continues on to other neurons. After generating an action potential, the cell enters a refractory period of several milliseconds, during which it returns to its resting potential in preparation for the generating of another pulse.

Chapter 3

The artificial neuron

3.1 Introduction

On the basis of the capacities and properties of the biological neuron, the first models for the artificial neurons were developed. These are the elements of the neural networks, which have properties designed to simulate the biological neurons. The connection topology between the artificial neurons is similar to the anatomy of the nervous system. Artificial neural networks have a lot of the same properties as the brain, for example the ability to learn on the basis of observations.
Let us now look at one of the first models for an artificial neural network.

3.2 The McCulloch-Pitts model

In 1943 McCulloch and Pitts published the first systematic study of artificial neural networks. They were the first to notice that the properties of the nervous system could give a model for a computer.
They made four assumptions, based on the neuro-picture of the time:

1. A neuron is either firing at time n or not.

2. A neuron cannot fire if at least one inhibitory synapse is active.

3. A neuron will fire if there are no active inhibitory synapses and if more than f excitatory synapses are active, where f is a fixed positive number.

4. The network topology is fixed for all time.

In spite of the fact that these assumptions were not all exactly true, this model was very important. McCulloch and Pitts were the first who modelled the computational capacities of the brain. This model led to better models in the future. Now let us encapsulate their assumptions into a concrete mathematical model:

Index the neurons by $1, 2, \cdots, N$. Assume that the neurons are capable of multiple synapses and also self-synapses. Fix $f \geq 1$.

Define $W = [w_{ij}]$, a $N \times N$ matrix, by:

w_{ij} = the number of synapses from neuron i to neuron j ,

if there are no inhibitory synapses between i and j.

$= -\infty$, otherwise.

Define $F : \{0,1\}^N \to \{0,1\}^N$ by:

$$F(x) = y$$

where

$$y_j = \begin{cases} 1 & \text{if } \sum_i x_i w_{ij} \geq f \Leftrightarrow x \cdot w_j \geq f \\ 0 & \text{otherwise} \end{cases}$$

where $w_j = (w_{1j}, w_{2j}, \cdots, w_{Nj})$ is the j^{th} column of W.

Here x is the input signal and y the output signal, defined as follows:

$$x_i = \begin{cases} 1 & \text{when neuron } i \text{ has sent a charge} \\ 0 & \text{otherwise} \end{cases}$$

and

$$y_j = \begin{cases} 1 & \text{when neuron } j \text{ will send a charge through its axon} \\ 0 & \text{otherwise} \end{cases}$$

In the fifties people developed another, more general model, based on the McCulloch-Pitts model.

This model is the following:

A neuron j will have synapses from other neurons. Each synapse is given a weight w_{ij}. The activity of the neuron which leads to this synapse has a value, which we call x_i.

The neuron also has a threshold θ_j.

The total activation of the neuron is: $s_j = -\theta_j + \sum_i w_{ij} x_i$.

The activity of the neuron is $f(s_j)$,

where f is a bounded measurable function, called the transfer function.

Based on this model F. Rosenblatt developed in 1958 a special neural network, called the Perceptron. We'll study this neural network in the next chapter.

Chapter 4

The perceptron

4.1 Introduction

The perceptron is a neural network invented by F.Rosenblatt and described in a paper published in 1958. He developed this network based on the ideas of McCulloch and Pitts. In 1962 Rosenblatt proved a theorem, dealing with the learning ability of the perceptron. This theorem showed that the perceptron can be trained to classify vectors into two groups. It does this automatically and always succeeds when possible, that is when the vectors can be linearly separated from each other.

In the 1960's perceptrons created a great deal of interest and optimism. The perceptrons were applied to such diverse problems as weather prediction, electrocardiogram analysis and artificial vision. It seemed for a time that the key to intelligence had been found; reproducing the human brain was only a matter of constructing a large enough network.

The initial euphoria was replaced by disillusionment as perceptrons were found to fail at certain simple learning tasks. The pereceptron failed to solve problems superficially similar to those they had been successful in solving. These unexplained failures launched a period of intense analysis. M.Minsky developed rigorous theorems regarding network operation. His research led in 1969 to the publication of the book "Perceptrons", in which he and S.Papert proved that the perceptrons were theoretically incapable of solving many simple problems. This book resulted into a near eclipse of the artificial neural network research. Discouraged researchers left

the field for areas of greater promise, government agencies redirected their funding and artificial neural networks lapsed into obscurity for nearly two decades.

The invention of the backpropagation algorithm in 1982 has played a large part in the resurgence of interest in artificial neural networks. Adding extra layers of neurons to the network led to a dramatic expanding of problems to which artificial neural networks can be applied. From then on there has been an explosive increase in the amount of research activity and better and better models have been developed to approximate the learning process of the brain.

Let us now look at the perceptron model, that has been developed in the early days of artificial neural networks.

4.2 The perceptron algorithm

The perceptron is a neural network composed of M processing units, which are also outputs of the system and N input units, where information from the environment is sent in. Each input unit n synapses onto each output unit m and a weight w_{nm} is assigned to this synapse. Each output unit m also has a threshold θ_m. These parameters are considered adjustable and will be changed by the system in response to experience(the "learning" of the network).

See figure 3.

An input vector $x = (x_1, \cdots, x_N) \in \mathbf{R}^{\mathrm{N}}$ is fed into the perceptron, where x_n is the activity of the input unit n. The output of the perceptron is a vector $y = y_w(x) = (y_1, \ldots, y_M) \in \{0,1\}^M$ where y_m is the activity of the output unit m and is defined by:

$$y_m = (y_w(x))_m = \begin{cases} 1 & \text{if } x \cdot w_m > -\theta_m \\ 0 & \text{otherwise} \end{cases}$$

where $w_m = (w_{1m}, w_{2m}, \ldots, w_{Nm}) \in \mathbf{R}^{\mathrm{N}}$.

Suppose that there are K non-zero vectors $x_1, \ldots, x_K \in \mathbf{R}^{\mathrm{N}}$.

There are also M properties, which each of these vectors either possess or not, that is:

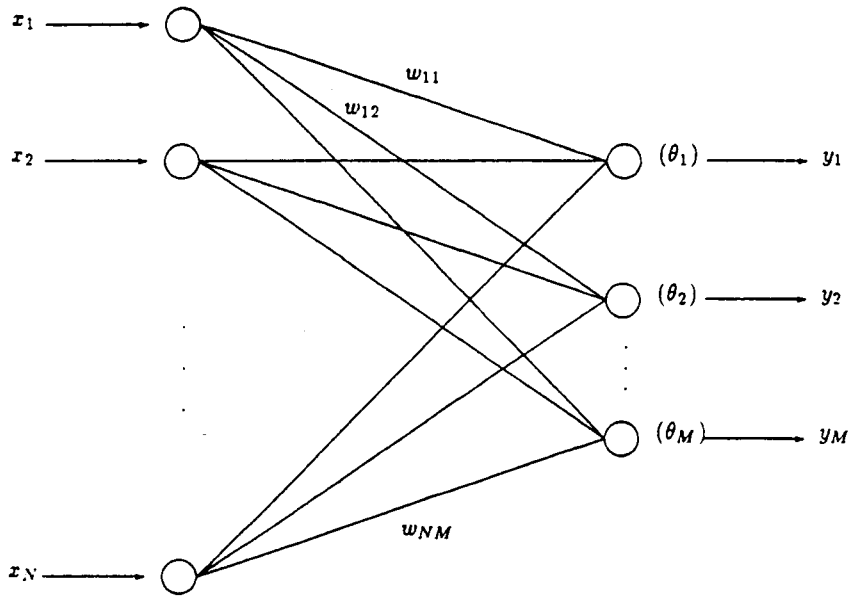

Figure 3. The perceptron

$\exists t : \mathbf{R}^{\mathrm{N}} \to \{0,1\}^{\mathrm{M}}$ such that $\forall i \in \{1,\ldots,K\} \forall m \in \{1,\ldots,M\}$:

$$(t(x_i))_m = \begin{cases} 1 & \text{if } x_i \text{ has property } m \\ 0 & \text{otherwise} \end{cases}$$

$t(x_i)$ Is called the desired output and $y_w(x_i)$ the observed output for input vector x_i and weight matrix W.

Now suppose that $X_1, X_2, \ldots$ is an arbitrary sequence, randomly chosen from $\{x_1,\ldots,x_K\}$. We say that the input vector X_i has been classified correctly if $t(X_i) = y_w(X_i)$, so when the desired output and the observed output are the same. After a correct classification we don't have a reason to adjust our weight matrix, but after a misclassification we do. We try to construct an algorithm in which the perceptron adjusts its weights to try to "learn" the classifications given by the function t.

That is we try to construct an algorithm which adjusts the weights such that after a finite number of misclassifications every input vector X_i will be correctly classified, i.e. we try to find a weight matrix W such that:

$$t(X_i) = y_w(X_i) \text{ for all input vectors } X_i.$$

We first make some reductions to simplify notation and that will lose no desired generality:

- Assume $M = 1$. This can be done without loss of generality since our learning rule for property m will only involve w_m, θ_m and the m^{th} coordinate of $t(X_i)$.

  The perceptron function now becomes:

  $$y_w(x) = 1_{(x \cdot w > -\theta)}.$$

- Next assume $\theta = 0$. This is also no loss since the effect of θ may be recaptured by adding an additional input unit whose activity is always 1:

  If $x = (x_1, \ldots, x_N)$ and $w = (w_1, \ldots, w_N) \in \mathbf{R}^N$ set $\tilde{x} = (x_1, \ldots, x_N, 1)$ and $\tilde{w} = (w_1, \ldots, w_N, \theta)$.

  Then there holds:

  $$y_w(x) = 1 \Leftrightarrow x \cdot w > -\theta \Leftrightarrow \tilde{x} \cdot \tilde{w} > 0.$$

So the **perceptron function** has become after these reductions:

$$y_w(x) = \begin{cases} 1 & \text{if } x \cdot w > 0 \\ 0 & \text{otherwise} \end{cases}$$

Now for the learning rule:

At time 0 begin with an arbitrary weight $w_0 \in \mathbf{R}^N$.
The weight at time n will be w_n.
At this time the perceptron receives the input vector X_{n+1}.
There are four possibilities:

1. $t(X_{n+1}) = y(X_{n+1}) = y_{w_n}(X_{n+1}) = 1$

   So X_{n+1} has been classified correctly, when the weight vector w_n is used. We have no reason to change our weight vector, so the algorithm becomes:

   $$w_{n+1} = w_n$$

2. $t(X_{n+1}) = y(X_{n+1}) = 0$

   Same story, so the algorithm is again:

$$w_{n+1} = w_n$$

3. $t(X_{n+1}) = 1$ and $y(X_{n+1}) = 0$

   So X_{n+1} hasn't been correctly classified when we use the weight vector w_n. We want to change the weight in such a way that the perceptron will be "closer" to giving a desired value of 1 the next time it has X_{n+1} as its input vector. So we want to make the innerproduct $X_{n+1} \cdot w$ larger. We define:

$$w_{n+1} = w_n + \eta X_{n+1} \text{ with } \eta > 0 \text{ fixed}$$

4. $t(X_{n+1}) = 0$ and $y(X_{n+1}) = 1$

   Same story as in 3, but now we want to make the innerproduct $X_{n+1} \cdot w$ smaller. We set:

$$w_{n+1} = w_n - \eta X_{n+1}$$

Combining these four cases, we derive the following **perceptron algorithm:**

$$w_{n+1} = w_n + \Delta w_n = w_n + \eta[t(X_{n+1}) - y_w(X_{n+1})]X_{n+1}$$

Here η is a fixed, positive parameter.
This process of adjustments of the weights results in the "training" of the network. The network is "learning" itself so that it can correctly classify the input vectors.

4.3 The perceptron convergence theorem

We will show that if a weight vector exists to correctly classify $x_1, \ldots, x_K$ then the above perceptron learning rule will converge. In particular there will be a N_0 such that:

$$n \geq N_0 \Rightarrow \Delta w_n = 0$$

First a definition.

Definition 4.1 (Linearly separability)

Let $F \subseteq \mathbf{R}^N \backslash \{0\}$.

Let $t : F \rightarrow \{0, 1\}$.

Define $\bar{F} = \{(x, t(x)) : x \in F\}$.

We say $\bar{F}$ *is linearly separable if there is a hyperplane in* $\mathbf{R}^N$ *so that all vectors* $x \in F$ *with* $t(x) = 1$ *are on one side and all vectors* $x \in F$ *with* $t(x) = 0$ *are on the other. That is*

if $\exists w \in \mathbf{R}^N$ *so:*

$$\begin{cases} x \cdot w > 0 \text{ for all } x \in F \text{ with } t(x) = 1 \\ x \cdot w \leq 0 \text{ for all } x \in F \text{ with } t(x) = 0 \end{cases}$$

Let $G = \{(x_k, t(x_k)) \ , 1 \leq k \leq K\}$, with $\{x_1, \ldots, x_K\}$ a collection of non-zero input vectors, as in section 4.2. Now define $\tilde{G}$ to be the following:

$$\tilde{G} = \{\tilde{x}_k\} \subseteq \mathbf{R}^N \setminus \{0\}, \text{with } \begin{cases} \tilde{x}_k = x_k & \text{if } t(x_k) = 1 \\ \tilde{x}_k = -x_k & \text{if } t(x_k) = 0 \end{cases}$$

There holds the following:

$$G \text{ is linearly separable } \Leftrightarrow \exists \tilde{w} \text{ such that } x \cdot \tilde{w} > 0 \ , \forall x \in \tilde{G}$$

So without loss of generality we can make our final reduction and assume that

$$t(x_k) = 1 \ , \forall 1 \leq k \leq K,$$

because the we consider $\tilde{G}$ instead of G.

Now let's formulate and prove the perceptron convergence theorem.

Theorem 4.1 (Perceptron Convergence Theorem, Rosenblatt 1962)

Let $F \subseteq \mathbf{R}^N$ *satisfy* $t(x) = 1, \forall x \in F$.

Let $X_1, X_2, \ldots$ *be an arbitrary sequence from* F.

Let w_0 *be arbitrary and* $\eta > 0$ *fixed. Let* w_n *be the weight vector at time* n.

Assume that F *is linearly separable, i.e.*

$$\exists \delta > 0 \ \exists \tilde{w} \in \mathbf{R}^N, \| \tilde{w} \| < \infty \text{ s.t. } \tilde{w} \cdot x \geq \delta > 0 \ , \forall x \in F$$

and also assume

$$\exists M < \infty \; so \; \| x \| \leq M \quad , \forall x \in F$$

Then the perceptron converges, i.e.

$$\exists N_0 \; so \; [\; w_n = w_{N_0} \; and \; y(X_n) = 1 \;] \quad , \forall n \geq N_0$$

So there will be only a finite number of misclassifications and a finite number of weight adjustments.

Proof

Suppose that there had been L_n changes in w at time n (that is L_n misclassifications) and that one of these just occurred, which is no restriction.

Since $\Delta w_n \neq 0$ we have:

$$\tilde{w} \cdot w_{n+1} \; = \; \tilde{w} \cdot (w_n + \Delta w_n) \; = \; (\tilde{w} \cdot w_n) + \eta(\tilde{w} \cdot X_{n+1}) \; \geq \; \tilde{w} \cdot w_n + \eta\delta$$

By applying the same procedure we find:

$$\tilde{w} \cdot w_{n+1} \; \geq \; \tilde{w} \cdot w_0 + \eta(L_n - 1)\delta + \eta\delta \; = \; \tilde{w} \cdot w_0 + \eta L_n \delta \tag{4.1}$$

Also:

$$\| w_{n+1} \|^2 = \; (w_n + \Delta w_n) \cdot (w_n + \Delta w_n) \; =$$

$$= \| w_n \|^2 + 2\eta(X_{n+1} \cdot w_n) + \eta^2 \| X_{n+1} \|^2$$

$$\leq \| w_n \|^2 + 2\eta(X_{n+1} \cdot w_n) + \eta^2 M^2$$

$$\leq \| w_n \|^2 + \eta^2 M^2$$

The last inequality follows because at time n we had a misclassification, so $y(X_{n+1})$ had to be 0 and so $X_{n+1} \cdot w_n < 0$.

Following the same procedure we find:

$$\| w_{n+1} \|^2 \leq \| w_0 \|^2 + (L_n - 1)\eta^2 M^2 + \eta^2 M^2 \; = \| w_0 \|^2 + L_n \eta^2 M^2 \tag{4.2}$$

Combining (4.1), (4.2) and using Cauchy-Schwarz, we conclude:

$$\tilde{w} \cdot w_0 + \eta L_n \delta \; \leq \; \tilde{w} \cdot w_{n+1} \; \leq \| \tilde{w} \| \, \| w_{n+1} \| \leq \| \tilde{w} \| \; \sqrt{\| w_0 \|^2 \; + L_n \eta^2 M^2}$$

Hence:

$$\tilde{w} \cdot w_0 + \eta L_n \delta \leq \| \tilde{w} \| \sqrt{\| w_0 \|^2 + L_n \eta^2 M^2}$$

So L_n is bounded in a rather implicit way, with a bound which doesn't depend on n. That finishes the proof.

□

Chapter 5

The case where the input vectors aren't linearly separable

5.1 Introduction

In chapter 4 we constructed an algorithm such that, in the case where the input vectors are bounded and can be linearly separated from each other, after a finite number of misclassifications every input vector X_i will be correctly classified. So if there exists a weight vector w that can linearly separate the input vectors from each other, then the perceptron algorithm converges to such a vector. But the linearly separability-assumption is a very strong assumption. In most cases in practice the two sets $\{x \ : \ t(x) = 1\}$ and $\{x \ : \ t(x) = 0\}$ aren't linearly separable. In this case we cannot tell from the perceptron convergence theorem whether our algorithm converges or not.

From now on we'll study classification problems where the input vectors aren't linearly separable.

5.2 A sufficient condition for ergodicity

Let the set of input vectors $\{Z_n \ , \ n \geq 0\}$ be an independent identically distributed (i.i.d.) sequence, defined on a space $\mathcal{Z} \subseteq \mathbf{R}^n$. Assume that $\mathrm{E}\| \ Z_1 \ \| < \infty$, i.e.

$$\int_{\mathbf{R}^n} \| \ z \ \| \ \mathrm{d}\nu(z) < \infty$$

where $\| \cdot \|$ denotes the Euclidian norm in $\mathbf{R}^n$ and $\nu(\cdot)$ is the induced probability measure of Z_1.

Also make the assumption that the measure ν is absolutely continuous w.r.t. Lebesgue measure λ on $\mathbf{R}^n$, that is

$$\forall A \subset \mathbf{R}^n \ : \ \lambda(\mathrm{A}) = 0 \text{ implies } \nu(\mathrm{A}) = 0$$

So Z_1 has a density function, denoted by f.

Finally make the same reduction as before, namely that $t(x) = 1$ for all input vectors x.

Remember the perceptron algorithm:

$$w_{n+1} = w_n + \eta Z_{n+1}[t(Z_{n+1}) - y_{w_n}(Z_{n+1})]$$

with η a fixed positive number and

$$y_w(Z_{n+1}) = 1_{(w_n \cdot Z_{n+1} > 0)}$$

In our case we can rewrite this algorithm in the following way:

$$w_{n+1} = w_n + \eta Z_{n+1} 1_{(w_n \cdot Z_{n+1} \leq 0)}$$

Assume that $w_0 \in \mathbf{R}^n$ is a random vector.

The process $\{w_k, k \geq 0\}$ is a Markov process, i.e. the conditional distribution of w_{n+1} given the values of $\{w_0, w_1, \ldots, w_n\}$ just depends on the value of w_n:

$$P(w_{n+1} \in A \mid w_0 = x_0, \ldots, w_n = x_n) \ = \ P(w_{n+1} \in A \mid w_n = x_n)$$

Knowing that our weight vectors form a Markov process, we can prove the following lemma:

Lemma 5.1

If we define

$$P(x, A) := P(w_{n+1} \in A \mid w_n = x) \ , \ x \in \mathbf{R}^n \ , \ \mathrm{A} \subseteq \mathbf{R}^n$$

then $\{ P(x, \cdot) \}$ *is strongly continuous, i.e.*

$\forall A \subseteq \mathbf{R}^n$, *Borel measurable* : $P(x, A)$ *is a continuous function in* x

Proof

Fix $A \subseteq \mathbf{R}^n$, a Borel measurable set. Fix $x \in \mathbf{R}^n$.

Then either $x \in \bar{A}$ or $x \notin \bar{A}$.

Case 1 : $x \notin \bar{A}$

Because $\bar{A}$ is closed we can find an $\hat{x}$, such that $\| x - \hat{x} \|$ is small enough so that $\hat{x} \notin \bar{A}$.

Now:

$$
\begin{aligned}
& P(x, A) \\
&= P(x + \eta\, Z\, 1_{(x \cdot Z \leq 0)} \in A) \\
&= P(x + \eta Z \in A\,,\ x \cdot Z \leq 0) + P(x \in A\,,\ x \cdot Z > 0) \\
&= P(x + \eta Z \in A\,,\ x \cdot Z \leq 0) \\
&=: P(B_x \cap C_x)
\end{aligned}
$$

where $B_x := \{x + \eta Z \in A\}$, $C_x := \{x \cdot Z \leq 0\}$.

The same calculation gives: $P(\hat{x}, A) = P(B_{\hat{x}} \cap C_{\hat{x}})$.

Now we have by the $\triangle$-inequality:

$$
\begin{aligned}
& | P(x, A) - P(\hat{x}, A) | \\
=& \\
& | P(B_x \cap C_x) - P(B_{\hat{x}} \cap C_{\hat{x}}) | \\
\leq& \\
& | P(B_x \cap C_x) - P(B_{\hat{x}} \cap C_x) | + | P(B_{\hat{x}} \cap C_x) - P(B_{\hat{x}} \cap C_{\hat{x}}) | \\
\leq& \\
& P(B_x \triangle B_{\hat{x}}) + P(C_x \triangle C_{\hat{x}}) \qquad (5.1)
\end{aligned}
$$

where $B_x \triangle B_{\hat{x}} = (B_x - B_{\hat{x}}) \cup (B_{\hat{x}} - B_x)$ denotes the symmetric difference.

Now

$$
\begin{aligned}
& P(B_x \triangle B_{\hat{x}}) \\
=& \\
& P(\{\eta Z + x \in A\} \triangle \{\eta Z + \hat{x} \in A\}) \\
=& \\
& P(\{\eta Z + x \in A\,,\ \eta Z + \hat{x} \notin A\} \cup \{\eta Z + \hat{x} \in A\,,\ \eta Z + x \notin A\})
\end{aligned}
$$

$=$

$$P(Z \in \tfrac{A-x}{\eta}\ ,\ Z \notin \tfrac{A-\hat{x}}{\eta}) \ +\ P(Z \in \tfrac{A-\hat{x}}{\eta}\ ,\ Z \notin \tfrac{A-x}{\eta}) \tag{5.2}$$

with

$$P(Z \in \tfrac{A-x}{\eta}\ ,\ Z \notin \tfrac{A-\hat{x}}{\eta})$$

$=$

$$\int 1_{\frac{A-x}{\eta}}(z)\, 1_{(\frac{A-\hat{x}}{\eta})^c}(z)\, f(z)\, \mathrm{d}z$$

$=$

$$\int 1_{\frac{A-x}{\eta}}(z)\, 1_{(\frac{A-x}{\eta}+\frac{x-\hat{x}}{\eta})^c}(z)\, f(z)\, \mathrm{d}z$$

$=$

$$\int 1_{(\frac{A-x}{\eta})^c}(z+\delta)\, 1_{\frac{A-x}{\eta}}(z)\, f(z)\, \mathrm{d}z \text{ , where } \delta := \tfrac{\hat{x}-x}{\eta}$$

$=:$

$$\int g(z+\delta)\, h(z)\, \mathrm{d}z$$

where the integrals are taken over $\mathbf{R}^n$ and $g(z) := 1_{(\frac{A-x}{\eta})^c}$ and $h(z) := 1_{(\frac{A-x}{\eta})}(z) f(z)$.

What happens if we let $\hat{x}$ converge to x, i.e. $\| \delta \| \to 0$?

$\lim_{\|\delta\|\to 0} g(z+\delta) \ =\ \lim_{\|\delta\|\to 0} 1_{(\frac{A-x}{\eta})^c}(z+\delta) \ =:\ \lim_{\|\delta\|\to 0} 1_B(z+\delta) \ =$

$= \lim_{\|\delta\|\to 0} 1_{(B-\delta)}(z) \ =\ 1_B(z) \ =\ g(z)$ a.e. $\mathbf{R}^n$

Also: $| g(z+\delta)h(z) | \leq f(z)$ for all z and δ, where f is an integrable function. So by the dominated convergence theorem we conclude that:

$$\lim_{\|\delta\|\to 0} \int g(z+\delta)\, h(z)\, \mathrm{d}z \ =\ \int g(z)\, h(z)\, \mathrm{d}z \ =\ \int 1_{(\frac{A-x}{\eta})^c}(z)\, 1_{(\frac{A-x}{\eta})}(z)\, f(z)\, \mathrm{d}z \ =\ 0$$

We get: $P(\ Z \in \frac{A-x}{\eta}\ ,\ Z \notin \frac{A-\hat{x}}{\eta}\ ) \longrightarrow 0$ as $\| \hat{x} - x \| \to 0$ (5.3)

The same calculation gives:

$$P(\ Z \in \frac{A-\hat{x}}{\eta}\ ,\ Z \notin \frac{A-x}{\eta}\ ) \longrightarrow 0 \text{ as } \| \hat{x} - x \| \to 0 \tag{5.4}$$

By combining (5.2), (5.3) and (5.4) we get that:

$$P(\ B_x \triangle B_{\hat{x}}\ ) \longrightarrow 0 \text{ as } \| \hat{x} - x \| \to 0 \tag{5.5}$$

Now look at:

$$C_x \triangle C_{\hat{x}} \ =\ \{\ Z\ :\ x \cdot Z \leq 0, \hat{x} \cdot Z > 0\ \cup\ x \cdot Z > 0, \hat{x} \cdot Z \leq 0\ \}$$

Then, as $\| \hat{x} - x \| \to 0$ this set converges to a (n-1)-dimensional hyperplane $H_x := \{Z : x \cdot Z = 0\}$, which has Lebesgue measure zero in $\mathbf{R}^n$.

For example see figure 4.

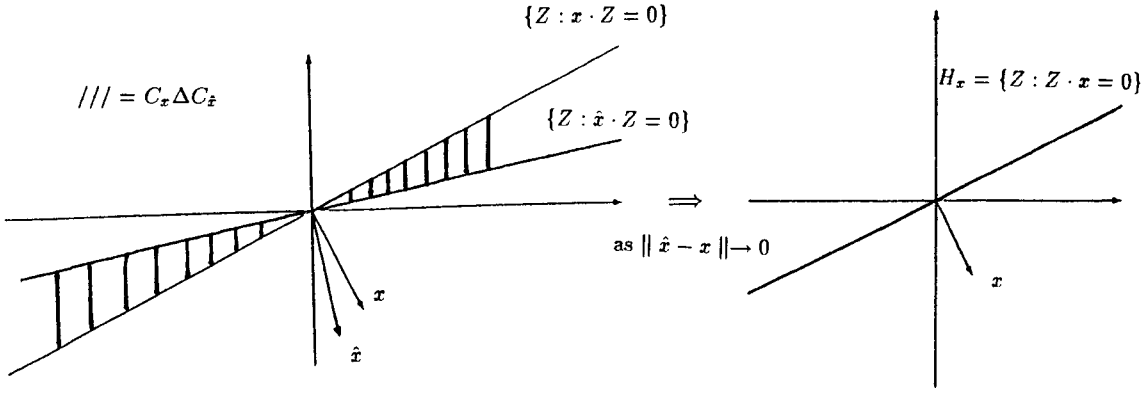

Figure 4. An illustration for case 1 in Lemma 5.1

We assumed that $\nu = PZ^{-1}$ is absolutely continuous w.r.t. the Lebesgue measure λ in $\mathbf{R}^n$, i.e.

$$\forall A \subset \mathbf{R}^n \ : \ \lambda(\mathrm{A}) = 0 \text{ implies } \nu(\mathrm{A}) = 0$$

Hence:

$$P(\ C_x \Delta C_{\hat{x}}\ ) \longrightarrow P(\ Z \in H_x\ ) \ = \ \nu(H_x) \ = 0 \text{ as } \ \| \hat{x} - x \| \to 0 \tag{5.6}$$

So by (5.1), (5.5) and (5.6) we conclude that:

$$|\ P(x, A) - P(\hat{x}, A)\ | \longrightarrow 0 \text{ as } \ \| \hat{x} - x \| \to 0$$

that is

$$P(x, A) \text{ is a continuous function in } x \text{ when } x \notin \bar{A}$$

Case 2 : $x \in \bar{A}$

Fix $\epsilon > 0$.

Take an open set $U \subset \mathbf{R}^n$ such that $x \in A_1 := U \cap \bar{A} \subset \bar{A}$ and $P(y, A_1) < \frac{\epsilon}{2}$ for all $y \in \mathbf{R}^n$.

Define $A_2 := \bar{A} \backslash A_1 \subset \bar{A}$.

Then:

$$P(x, A)$$

$$=$$

$$P(\, x + \eta Z 1(x \cdot Z \leq 0) \in A \,)$$

$$=$$

$$P(\, x + \eta Z 1(x \cdot Z \leq 0) \in A_1 \,) \; + \; P(\, x + \eta Z 1(x \cdot Z \leq 0) \in A_2 \,)$$

$$=$$

$$P(x, A_1) \; + \; P(x, A_2)$$

Now:

$$\mid P(x, A) \; - \; P(\hat{x}, A) \mid$$

$$=$$

$$\mid P(x, A_1) \; - \; P(\hat{x}, A_1) \; + \; P(x, A_2) \; - \; P(\hat{x}, A_2) \mid$$

$$\leq$$

$$\mid P(x, A_1) \; - \; P(\hat{x}, A_1) \mid \; + \; \mid P(x, A_2) \; - \; P(\hat{x}, A_2) \mid$$

$$<$$

$$\epsilon + \mid P(x, A_2) \; - \; P(\hat{x}, A_2) \mid$$

Now because $x \notin \bar{A}_2$, we can apply the results derived in case 1 for x and A_2 and conclude that:

$$\exists \delta_\epsilon \; : \parallel \; x - \hat{x} \; \parallel < \; \delta_\epsilon \implies \mid \; P(x, A_2) \; - \; P(\hat{x}, A_2) \; \mid < \; \epsilon$$

So we get:

$$\exists \delta_\epsilon \; : \parallel \; x - \hat{x} \; \parallel < \; \delta_\epsilon \implies \mid \; P(x, A) \; - \; P(\hat{x}, A) \; \mid < \; 2\epsilon$$

Hence:

$$P(x, A) \text{ is a continuous function in } x, \text{ when } x \in \bar{A}$$

We conclude, by combining case 1 and case 2, that

$$\forall A \subseteq \mathbf{R}^n \; : \; \mathrm{P(x, A)} \text{ is a continuous function in x}$$

□

Assume that the conditions preceding lemma 5.1 are still satisfied.

Also assume that we are in the non-separable case, that is no $w \in \mathbf{R}^n$ correctly classifies (this is called "frustration") i.e.

$$P(\, Z \cdot w \leq 0\,) \; > \; 0 \text{ for all } w \in \mathbf{R}^n$$

Then we have the following lemma:

Lemma 5.2

There exists a compact set $K \subset \mathbf{R}^n$, *with* $\lambda(K) > 0$, *such that:*

$\exists 0 < c < \infty$ *such that* $E(\, \| \, w_{n+1} \, \| - \| \, w_n \, \| \;\mid\; w_n = x\,) \leq -c \quad$ *for all* $x \notin K$

and

$\exists 0 < B < \infty$ *such that* $E(\, \| \, w_{n+1} \, \| - \| \, w_n \, \| \;\mid\; w_n = x\,) \leq B \quad$ *for all* $x \in K$

Proof

Define $A_x := P(\, Z \cdot x \leq 0\,)$: the probability that a misclassification occurs at $x \neq 0$.

Then $A_{dx} = P(\, Z \cdot dx \leq 0\,) = P(\, Z \cdot x \leq 0\,) = A_x$ for all $d \in (0, \infty)$.

So in studying the function $x \rightarrow A_x$ we can restrict to the case where

$x \in S^1 = \{x \in \mathbf{R}^n \mid \| \, x \, \| = 1\}$.

Remember the definition of $C_x := \{\, x \cdot Z \leq 0\}$.

Take a $\hat{x}$ close to x, then:

$$\mid A_x - A_{\hat{x}} \mid$$

$=$

$$\mid P(Z \cdot x \leq 0) \; - \; P(Z \cdot \hat{x} \leq 0) \mid$$

$=$

$$\mid P(C_x) \; - \; P(C_{\hat{x}}) \mid$$

$\leq$

$$P(\, C_x \triangle C_{\hat{x}}\,)$$

So by (5.6) we conclude that:

$$\mid A_x \; - \; A_{\hat{x}} \mid \longrightarrow 0 \text{ as } \| \; \hat{x} \; - \; x \; \| \rightarrow 0$$

i.e. $x \to A_x$ is continuous on S^1.

By our assumption of frustration we also have that $A_x > 0$ for all $x \in S^1$.

So by compactness of S^1 : $\exists \gamma > 0$ s.t. $A_x \geq \gamma > 0 \; \forall x \in S^1$

Fix $x \in S^1$.

Define $L(x) \equiv \{y : x \cdot y \leq 0\}$.

Suppose $y \in L(x)$, then for $k \geq 1$:

$$\| kx + \eta y \|^2 = \| kx \|^2 + \| \eta y \|^2 + 2k\eta(x \cdot y)$$

So:

$$(\| kx + \eta y \| - \| kx \|)(\| kx + \eta y \| + \| kx \|) = \| \eta y \|^2 + 2k\eta(x \cdot y)$$

Hence:

$$\| kx + \eta y \| - \| kx \| = \frac{\| \eta y \|^2 + 2k\eta(x \cdot y)}{\| kx + \eta y \| + \| kx \|} = \frac{\frac{1}{k} \| \eta y \|^2 + 2\eta(x \cdot y)}{\| x + \frac{\eta}{k} y \| + \| x \|}$$

So:

$$\| kx + \eta y \| - \| kx \| \longrightarrow \frac{2\eta(x \cdot y)}{2 \| x \|} = \eta(x \cdot y) \quad \text{as } k \to \infty$$

Also, by the Δ- inequality:

$$\| kx + \eta y \| - \| kx \| \leq \| \eta y \|$$

with $\int \| \eta y \| \, d\nu(y) = \eta \int \| y \| \, d\nu(y) < \infty$, by the assumption we made in the beginning of this section.

So by the Dominated Convergence Theorem we get

$$lim_{k \to \infty} \int_{L(x)} [\, \| kx + \eta y \| - \| kx \| \,] \, d\nu(y) = \eta \int_{L(x)} (x \cdot y) \, d\nu(y) < 0$$

by definition of $L(x)$ and our frustration-assumption.

Hence:

$$\forall x \; \exists k_x \; \forall k \geq k_x \; : \; \int_{L(x)} \| kx + \eta y \| - \| kx \| \; d\nu(y) < 0$$

The function $x \to k_x$ is continuous.

For $\hat{x} \in S^1$ close enough to x, there exists $c_x > 0$ such that

$$\int_{L(\hat{x})} \| k_x \hat{x} + \eta y \| - \| k_x \hat{x} \| \; d\nu(y) < -c_x < 0$$

So $\forall k \geq k_x \, \forall \hat{x}$ in a small open neighborhood N_x of x :

$$\int_{L(\hat{x})} \parallel k\hat{x} + \eta y \parallel - \parallel k\hat{x} \parallel \; \mathrm{d}\nu(y) \; < -c_x < 0$$

The union of $\{N_x : x \in S^1\}$ forms an open cover of S^1. Then by compactness of S^1 we can find a set $\{x_1, \ldots, x_L\}$ such that their corresponding neighborhoods $N_{x_1}, \ldots, N_{x_L}$ form a finite open subcover of S^1.

Now define

$$k_0 \equiv \max(k_{x_1}, \ldots, k_{x_L})$$

and

$$c \equiv \min(c_{x_1}, \ldots, c_{x_L})$$

Then we can conclude

$$\forall k > k_0 \forall x \in S^1 \; : \; \int_{L(x)} [\parallel kx + \eta y \parallel - \parallel kx \parallel] \, \mathrm{d}\nu(y) \leq -c < 0$$

So:

$$\forall k > k_0 \forall x \in S^1 \; : \; \mathrm{E}(\parallel w_{n+1} \parallel - \parallel kx \parallel \mid w_n = kx) \leq -c$$

Define $K := \{x \mid \parallel x \parallel \leq k_0\}$, a compact set, with $\lambda(K) > 0$. Then:

$$E(\parallel w_{n+1} \parallel - \parallel x \parallel \mid w_n = x) \leq -c \text{ for all } x \notin K$$

Observe that $x \rightarrow E(\parallel w_{n+1} \parallel \mid w_n = x)$ is continuous,

so $E(\parallel w_{n+1} \parallel - \parallel x \parallel \mid w_n = x)$ is a continuous function of x.

Then by compactness of K we conclude:

$$\exists B > 0 \; : \; E(\parallel w_{n+1} \parallel - \parallel x \parallel \mid w_n = x) \leq B < \infty \text{ for all } x \in K$$

because a continuous function on a compact set is bounded.

□

The results of Lemma 5.1 and Lemma 5.2 imply that, under the assumptions we made, the conditions in Theorem 5.1 [Tweedie,1975] are satisfied. These conditions are sufficient conditions for our Markov process to be ergodic. So we can apply this theorem and conclude the following:

Theorem 5.1

Define the following Markov process in $\mathbf{R}^n$*:*

$$w_{n+1} = w_n + \eta Z_{n+1} 1(w_n \cdot Z_{n+1} \leq 0)$$

with w_0 *arbitrary and* $\eta > 0$ *fixed.*

Assume $\{Z_n\ ,\ n \geq 0\}$ *is an i.i.d. sequence, defined on a space* $\mathcal{Z} \subseteq \mathbf{R}^n$*, with* $E\| Z_1 \| < \infty$*. Assume that the induced probability measure* ν *of* Z_n *is absolutely continuous w.r.t. Lebesgue measure* λ*.*

Finally assume non-separability:

$$P(\ Z \cdot w \leq 0\ ) > 0\ ,\ \forall w \in \mathbf{R}^n\ ,\ \forall \mathrm{Z} \in \mathcal{Z}$$

Then the sequence $\{w_n\ ,\ n \geq 0\}$ *is an ergodic Markov process,*

that is there exists an unique invariant probability distribution μ^η*, i.e.*

$$\exists!\ \mu^\eta\ s.t.\ \mu^\eta(A) = \int P(y, A)\, \mathrm{d}\mu^\eta(y)\ ,\ \forall A \in \mathcal{B}(\mathbf{R}^n).$$

We have the following corollary, as a consequence of theorem 5.1 :

Corollary 5.1 (Tweedie,1974)

Suppose that we have the same assumptions as in Theorem 5.1.

Then:

$$\frac{1}{n}\sum_{i=1}^{n} P(w_{k+i} \in A \mid w_k = y) \xrightarrow{a.s.} \mu^\eta(A)\ as\ n \to \infty.$$

Actually we can replace this corollary by the following stronger statement:

Corollary 5.2 (Pollard and Tweedie,1975)

Assume that the conditions in theorem 5.1 hold.

Then for any initial distribution ν *on* $\mathbf{R}^n$ *we have:*

$$\| \frac{1}{n} \int \sum_{i=1}^{n} P^i(y, \cdot)\mathrm{d}\nu(y) - \mu^\eta(\cdot) \|_{TV} \longrightarrow 0\ as\ n \to \infty$$

where $\| \cdot \|_{TV}$ *denotes the total variation norm and* $P^i(y, x) = P(w_{n+i} = x \mid w_n = y)$*.*

A special application, where the conditions in the theorem are satisfied, is the case where the inputvectors $\{Z_n, n \geq 0\}$ are normally distributed.

We'll study this case in the next chapter.

Chapter 6

The Gaussian case

6.1 Introduction

Consider again the perceptron algorithm:

$$w_{n+1} = w_n + \eta X_{n+1} 1_{(w_n \cdot X_{n+1} \leq 0)}$$

In chapter 5 we saw sufficient conditions for this weight process to be ergodic in the non-separable case. We are interested in non-separable cases where we have geometric ergodicity for this weight process, that is:

$$P(w_{k+n} \in A \mid w_k = y) \longrightarrow \mu^\eta(A) \quad \forall A \, \forall y \text{ , as } n \to \infty$$

where μ^η denotes again our unique invariant distribution.
One example, where the conditions in theorem 5.1 are satisfied is the case where the input vectors are normally distributed. So applying this theorem we conclude that in the Gaussian case our weight process is ergodic. We'll see in this chapter that a slightly modificated weight process is geometric ergodic in the two-dimensional Gaussian case, so that in this situation the process converges in distribution to the invariant distribution.

From now on assume that the input vectors and the weights are two-dimensional vectors.
The aim of the perceptron algorithm is to construct a vector w, such that the separation line $\{x \ : \ x \cdot w = 0\}$ results in a minimal number of classification errors.

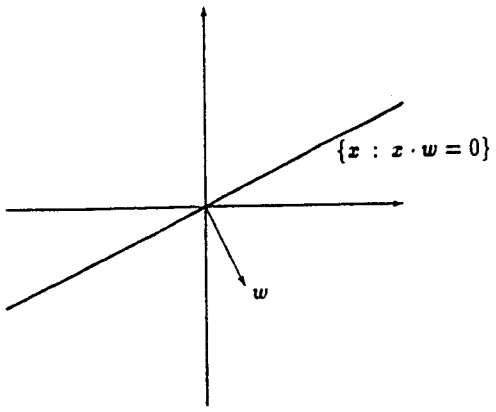

Figure 5. The two-dimensional case

In the separable case we concluded from the perceptron convergence theorem that $w_n \to w^*$, where the line $\{x : x \cdot w^* = 0\}$ is such that we don't make any classification errors.

In the construction of the optimal separation line the length of the vector w is of no importance. We are just interested in the angle that w makes with the x-axis, because this angle determines the direction of the separation line. It doesn't matter for the separation line if we make our vector w longer or shorter to a vector with the same direction and with length 1.

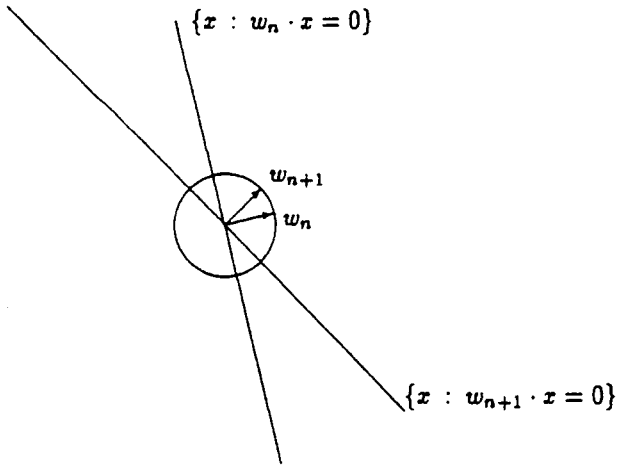

Figure 6. The circle S^1

That's why, without loss of generality, we can restrict to the case where $\| w \| = 1$ for our weight vectors. So from now on we work on the compact set S^1, instead of the space $\mathbf{R}^2$.

We change the algorithm as follows:

- w_0 arbitrary, $\| w_0 \| = 1$

(6.1)

- $w_{n+1} = \frac{w_n + \eta(t(X_{n+1}) - y_{w_n}(X_{n+1}))X_{n+1}}{\|w_n + \eta(t(X_{n+1}) - y_{w_n}(X_{n+1}))X_{n+1}\|}$

The angle between w_{n+1} and the x-axis stays the same, the length of w_{n+1} becomes 1. The separation line $\{x : x \cdot w_{n+1} = 0\}$ doesn't change. See figure 6.

In the next section we'll prove that this weight process, which is a Markov process on the circle S^1, is a geometric ergodic process when the input vectors are normally distributed.

6.2 Geometric ergodicity

Definition 6.1

Let F_{S^1} be the collection of Borel sets on S^1.

A function $p(\cdot, \cdot) : S^1 \times F_{S^1} \to [0, 1]$ is called a stochastic transition function if it satisfies the following properties:

1. *$p(\xi, \cdot)$ determines, for a fixed $\xi \in S^1$, a probability measure on F_{S^1}*
2. *$p(\cdot, A)$ determines, for a fixed $A \in F_{S^1}$, a F_{S^1}- measurable function*

The n-step transition probabilities, examples of stochastic transition functions, are calculated in the following inductive way:

- $p^{(1)}(\xi, A) = p(\xi, A) = P(w_{k+1} \in A \mid w_k = \xi)$
- $p^{(n+1)}(\xi, A) = P(w_{k+n+1} \in A \mid w_k = \xi) = \int_{S^1} p^{(n)}(\alpha, A) p(\xi, \mathrm{d}\alpha)$

The probability that w_n belongs to the set A is calculated in the following inductive way:

- $P(w_n \in A) =: p(A)$, for $n = 0$
- $P(w_n \in A) = \int_{S^1} p^{(n)}(\alpha, A)p(\mathrm{d}\alpha)$, for $n \geq 1$

When this probability doesn't depend on n, then we call the $\{w_n\}$-process strictly stationary and $p(\cdot)$ a stationary or invariant probability distribution.

$p^{(n)}(\xi, \cdot)$ is a measure on F_{S^1} for all ξ and n, therefore by the Lebesgue decomposition it has an absolutely continuous and a singular component with respect to Lebesgue measure φ on S^1, that is we can write:

$$p^{(n)}(\xi, A) = \int_A p_0^{(n)}(\xi, \alpha)\varphi(\mathrm{d}\alpha) + \Delta^{(n)}(\xi, A),$$

with $p_0^{(n)}(\xi, \cdot)$ a F_{S^1}-measurable function for all $\xi \in S^1$ and $\Delta^{(n)}(\xi, \cdot)$ a measure on F_{S^1}.

Theorem 6.1 (Doob,1953)

Let φ denote the Lebesgue measure on S^1.

Suppose:

$$\exists 0 < \delta < 1 \ s.t. \ p_0^{(1)}(\xi, \alpha) \geq \delta \ , \forall \xi, \alpha \in S^1 \qquad (6.2)$$

Then there exists an invariant probability distribution $\mu(\cdot)$ such that:

$$\mid p^{(n)}(\xi, A) - \mu(A) \mid \leq (1 - \delta)^{n-1} \ , n = 1, 2, \ldots$$

Can we apply this theorem to the weight process generated by (6.1)?
For this we have to determine whether condition (6.2) holds for this process.
Reduce again to the case where $t(x)$, the desired output of vector x, is identically equal to 1.

Theorem 6.2

Suppose the Markov process $\{w_n, n \geq 0\}$ *is generated by the following algorithm:*

- w_0 *random,* $\| w_0 \| = 1$

(6.3)

- $w_{n+1} = \frac{w_n + \eta(1_{(w_n \cdot X_{n+1}) \leq 0}) X_{n+1}}{\| w_n + \eta(1_{(w_n \cdot X_{n+1}) \leq 0}) X_{n+1} \|}$

where the inputvectors X_{n+1} *are normally* $N_2(\mu, I)$ *distributed.*

Define $p_0^{(1)}(\cdot,\cdot)$ *to be the absolutely continuous component of* $p^{(1)}(\cdot,\cdot)$*, the 1-step transition probabilities for* $\{w_n, n \geq 0\}$*, with respect to* φ*.*

Then $p_0^{(1)}(\cdot,\cdot)$ *satisfies condition (6.2).*

Proof

Let η, the learning parameter, be a fixed, positive number.

We are interested in the conditional distribution of w_{n+1}, given w_n.

The following holds:

$$P(w_{n+1} \in A \mid w_n = \xi) = P(f(w_n, X_{n+1}) \in A \mid w_n = \xi) = P(f(\xi, X_{n+1}) \in A)$$

because X_{n+1} is independent of w_n.

Here $A \subset S^1$ and

$$f(\xi, X_{n+1}) = \frac{\xi + \eta(1_{(\xi \cdot X_{n+1}) \leq 0}) X_{n+1}}{\| \xi + \eta(1_{(\xi \cdot X_{n+1}) \leq 0}) X_{n+1} \|} \tag{6.4}$$

So:

$$f(\xi, X_{n+1}) = \begin{cases} \xi & \text{with probability } p \equiv P(\xi \cdot X_{n+1} > 0) \\ \frac{\xi + \eta X_{n+1}}{\| \xi + \eta X_{n+1} \|} & \text{with probability } 1 - p \end{cases}$$

We try to calculate the distribution of $\frac{\xi + \eta X_{n+1}}{\| \xi + \eta X_{n+1} \|}$.

Define $Y_{n+1} \equiv \xi + \eta X_{n+1}$. Consider two cases:

Step 1. The calculation of $\mathcal{L}(Y_{n+1})$, the distribution of Y_{n+1}

We have $X_{n+1} \sim N_2(\mu, I)$, so $Y_{n+1} \sim N_2(\xi + \eta\mu, \eta^2 I)$.

The density of Y_{n+1} is then:

$$p(y) = \frac{1}{2\pi\eta} e^{-\frac{1}{2\eta^2}(y - \xi - \eta\mu)^T (y - \xi - \eta\mu)}$$

Step 2. The calculation of $\mathcal{L}(\frac{Y_{n+1}}{\|Y_{n+1}\|})$

$\frac{Y_{n+1}}{\|Y_{n+1}\|}$ Is a point on S^1, with coordinates (cosφ,sinφ).
Define the density of (cosφ,sinφ) on S^1, by the density of φ on $[0, 2\pi]$.

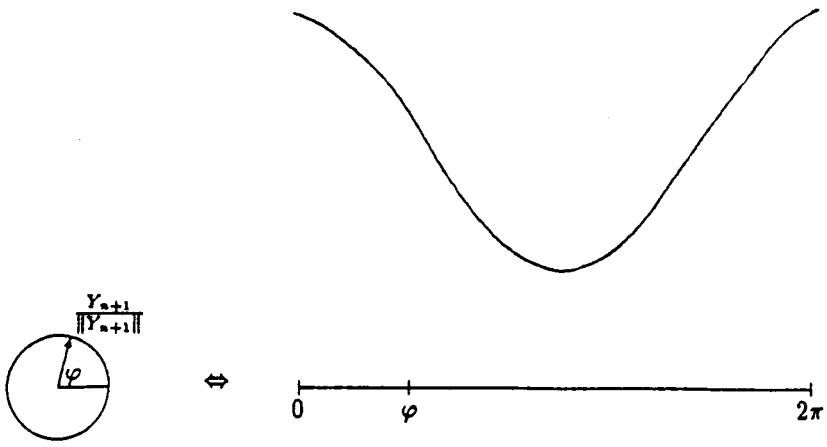

Figure 7. An example of a density function on $[0, 2\pi]$

Consider the following function:

$$g : (y_1, y_2) \rightarrow (\sqrt{y_1^2 + y_2^2}\,, \arctan(\frac{y_2}{y_1}))$$

or:

$$g : (r\cos\varphi, r\sin\varphi) \rightarrow (r, \varphi)$$

Define $q(\varphi)$ to be the density function of $\varphi = \arctan(\frac{y_2}{y_1})$. We are trying to calculate $q(\varphi)$. First calculate the density function of $(\sqrt{y_1^2 + y_2^2}, \arctan(\frac{y_2}{y_1}))$. Denote this function by $s(r, \varphi)$.

Then:

$s(r, \varphi)$

=

$p_{g(y_1,y_2)}(r, \varphi)$

=

$p_{(y_1,y_2)}(g^{-1}(r, \varphi))\,.\ |\det \mathrm{Jac}_{g^{-1}}(r, \varphi)\,|$

$=$

$p_{(y_1,y_2)}(r\cos\varphi, r\sin\varphi)\ .\ r$

$=$

$\frac{r}{2\pi\eta} e^{-\frac{1}{2\eta^2}(r\cos\varphi-\xi_1-\eta\mu_1\ ,\ r\sin\varphi-\xi_2-\eta\mu_2)(r\cos\varphi-\xi_1-\eta\mu_1, r\sin\varphi-\xi_2-\eta\mu_2)^T}$

Then:

$$q(\varphi) = \int_0^\infty s(r,\varphi)\mathrm{d}r$$

This is the density function of $\frac{Y_{n+1}}{\|Y_{n+1}\|}$.

There holds $s(r,\varphi) = r\ p(r\cos\varphi, r\ \sin\varphi) > 0\ ,\ \forall r > 0, \forall\varphi \in [0,2\pi]$, because p is a normal density function.

Hence:

$$q(\varphi) = \int_0^\infty s(r,\varphi)\mathrm{d}r > 0\ ,\ \forall\varphi \in [0,2\pi] \tag{6.5}$$

Also q is continuous, because:

- $r \rightarrow s(r,\varphi)$ is integrable over $[0,\infty)\ ,\ \forall\varphi \in [0,2\pi]$

  because $s(r,\varphi) = r\ p(r\cos\varphi, r\sin\varphi)$, with p a normal distribution

- $\varphi \rightarrow s(r,\varphi)$ is continuous in φ_0 for almost all $r \in [0,\infty)$

  because $\cos(\varphi)$ and $\sin(\varphi)$ are continuous functions,so

  $$\lim_{h\rightarrow 0} p(r\cos(\varphi+h), r\sin(\varphi+h)) = p(r\cos\varphi, r\sin\varphi)$$

  because p is a normal distribution, so it is continuous in both variables

- $\exists g(r)$ integrable, non-negative and independent of φ such that:

  $$|\ s(r,\varphi)\ | \leq g(r)\ ,\ \forall r \geq 0, \forall\varphi \in [0,2\pi]$$

  Intuitively it's clear that such a function g exists, because if $r \rightarrow \infty$ then $p(r\cos\varphi, r\sin\varphi) \rightarrow 0\ ,\ \forall\varphi \in [0,2\pi]$ and so $s(r,\varphi) \rightarrow 0\ ,\ \forall\varphi \in [0,2\pi]$.

  However it's hard to determine such a function g explicitely.

Now we can apply the lemma, an application of the dominated convergence theorem, which states that the three upper conditions imply that $q(\varphi) = \int_0^\infty s(r,\varphi)\mathrm{d}r$ is continuous on $[0,2\pi]$.

The continuity of q and equation (6.5) now imply that:

$$\exists \delta_0 > 0 \text{ such that } q(\varphi) > \delta_0 \ , \ \forall \varphi \in [0, 2\pi] \tag{6.6}$$

We conclude from (6.4) that:

$p^{(1)}(\xi, A)$

$=$

$P(w_{n+1} \in A \mid w_n = \xi)$

$= \{w_n \text{ and } X_{n+1} \text{ are independent}\}$

$P(\frac{\xi+\eta X_{n+1}}{\|\xi+\eta X_{n+1}\|} \in A \ , \ \xi X_{n+1} \leq 0) \ + \ P(\xi \in A \ , \ \xi X_{n+1} > 0)$

$= \{\xi \text{ and } X_{n+1} \text{ are independent}\}$

$P(\frac{\xi+\eta X_{n+1}}{\|\xi+\eta X_{n+1}\|} \in A \mid \xi X_{n+1} \leq 0) \, P(\xi X_{n+1} \leq 0) \ + \ P(\xi \in A) \, P(\xi X_{n+1} > 0)$

$=$

$(1-p) P(\frac{\xi+\eta X_{n+1}}{\|\xi+\eta X_{n+1}\|} \in A \mid \xi X_{n+1} \leq 0) \ + \ p \, P(\xi \in A)$

$= \{\xi \text{ is a fixed vector in } S^1\}$

$(1-p) \int_A q(\varphi) 1_{(\xi_1 r \cos\varphi + \xi_2 r \sin\varphi \leq 0)} \mathrm{d}\varphi \ + \ p\delta_\xi(A)$

with q the density of $\frac{Y_{n+1}}{\|Y_{n+1}\|}$ and $\delta_\xi(A)$ the Dirac measure of ξ on $A \subset S^1$.

Define $p_0^{(1)}(\xi, \alpha) \equiv (1-p) q(\varphi) 1_{(\xi_1 r \cos\varphi + \xi_2 r \sin\varphi \leq 0)}$ with $\alpha \equiv (r, \varphi)$.

Further define $\Delta^{(1)}(\xi, A) \equiv p\delta_\xi(A)$, then we get the following:

$$p^{(1)}(\xi, A) \ = \ \int_A p_0^{(1)}(\xi, \alpha) \varphi(\mathrm{d}\alpha) \ + \ \Delta^{(1)}(\xi, A)$$

with φ the Lebesgue measure on S^1.

From (6.6) we conclude:

$p_0^{(1)}(\xi, \alpha)$

$= (1-p) q(\varphi) 1_{(\xi_1 r \cos\varphi + \xi_2 r \sin\varphi \leq 0)}$

$\geq (1-p) q(\varphi)$

$\geq (1-p)\delta_0 \equiv \delta \quad , \ \forall \xi, \alpha \in S^1$

Hence we proved that $p_0^{(1)}(\xi, \alpha)$ satisfies (6.2)

□

Theorem 6.2 tells us that, in the case of two-dimensional Gaussian input vectors, the normalized weight sequence results in an absolutely continuous component of its 1-step transition probabilities, that satisfies (6.2).

Hence we can apply Theorem 6.1 and conclude:

Corollary 6.1

Suppose that the Markov process on S^1 *is generated by algorithm (6.3).*

Suppose that the input vectors X_n *are normally* $N_2(\mu, I)$ *distributed.*

Denote $p^{(n)}(\xi, A) = P(w_n \in A \mid w_0 = \xi)$.

Then there exists $0 < \delta < 1$ *and an invariant distribution* $\mu(\cdot)$ *such that:*

$$\mid p^{(n)}(\xi, A) - \mu(A) \mid \leq (1-\delta)^{n-1} \ , \ \forall n \geq 1 \ , \ \forall A \subset F_{S^1} \ , \ \forall \xi \in S^1$$

i.e.

$$p^{(n)}(\xi, A) \to \mu(A) \quad as\ n \to \infty \ , \ \forall A \subset F_{S^1} \ , \ \forall \xi \in S^1$$

6.3 Simulations

Consider our original perceptron algorithm:

$$w_{n+1} = w_n + \eta(t(X_{n+1}) - y_{w_n}(X_{n+1}))X_{n+1}$$

where:

$$y_{w_n}(X_{n+1}) = \begin{cases} 1 & \text{if } w_n \cdot X_{n+1} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Assume that the input vectors x with $t(x) = 1$ are $N_2((\mu, 0), I)$- distributed and the vectors with $t(x) = 0$: $N_2((-\mu, 0), I)$, with $\mu > 0$.

The optimal separation line, which results in a minimal number of expected classification errors is the y-axis. The second coordinate of the optimal weight vector is 0.

Let $\eta = 0.01$ and observe how the weight process developes in 2000 iterations.

Figure 9 (a) shows that the weight process in $\mathbf{R}^2$ converges to a neighborhood of the origin, when $w_0 = (0, 1)$. Figures 9 (b) and (c) show the same phenomena when the

starting weight vector is respectively $(0,-1)$ and $(-1,0)$. In figure 10 we see that when we start in $(1,0)$, we move around the x-axis to a neighborhood of the origin. These four processes occur under the influence of the following vector field:

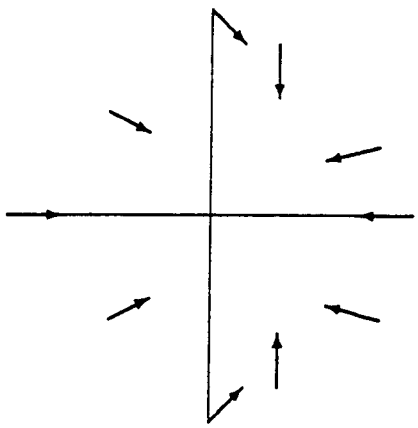

Figure 8. The vector field

Our main interest is the size of $\varphi_n = \arctan(\frac{w_n(2)}{w_n(1)})$, the angle between w_n and the x-axis. This is because the separation line $\{x : x \cdot w_n = 0\}$ is only depending on the angle φ_n and not on the length of the vector w_n. The optimal angle is in our case 0.

Figure 11 shows the graph of $n \rightarrow \frac{w_n(2)}{w_n(1)} = \tan(\varphi_n)$, for the case where $\eta = 0.01, N = 2000$ and $w_0 = (1,0)$. We see that until about 1100 iterations the angle remains close to the optimal angle 0. It describes the same process, that we saw in figure 10. However after 1100 iterations we are stuck in a small neighborhood of the origin and the process is oscillating under the influence of the discontinuity point $(0,0)$. The angle is fluctuating and doesn't stay in a neighborhood of 0.

In figure 12 we see the same phenomenon for $\eta = 0.002, N = 10,000$ and $w_0 = (1,0)$. After 6000 iterations the process is going to oscillate.

Now consider again the case where we have normalized weight vectors.
That is we consider the following algorithm:

$$w_{n+1} = \frac{w_n + \eta(t(X_{n+1}) - y_{w_n}(X_{n+1}))X_{n+1}}{\| w_n + \eta(t(X_{n+1}) - y_{w_n}(X_{n+1}))X_{n+1} \|}$$

Again assume that the input vectors x with $t(x) = 1$ are $N_2((\mu, 0), I)$-distributed and the vectors with $t(x) = 0 : N_2((-\mu, 0), I)$, where $\mu > 0$.

The optimal weight vector is $(1, 0)$, because for this vector the separation line is the y-axis and the input vectors with positive x- coordinate are classified to be 1, and the vectors with negative coordinate to be 0. Using the results from corollary 6.1 we know that the conditional distribution of w_n, given the value of w_0, converges to the invariant distribution $\mu = \mu^\eta$.

Assume that $w_0 = (1, 0)$. This is no restriction because in using the algorithm we could, after one step, end up in any other point on the circle with positive probability. Now if the function $\eta \to \mu^\eta$ is continuous, then this would imply that

$$\mu^\eta \to \mu^0 \quad , \text{ if } \eta \to 0$$

Now $\mu^0 = \delta_{(1,0)}$ because if $\eta = 0$ then after applying the algorithm the weights stay to be $(1, 0)$. So if $\eta \to \mu^\eta$ is continuous then we have convergence of the weight vectors to the optimal vector $(1, 0)$, if we let $\eta \to 0$. The question remains whether this assumption of continuity is true. We conjecture that the distribution of μ^η converges to the distribution concentrated at the optimal weight vector $(1, 0)$. This is actually supported by computer simulations, but we haven't been able to find a rigorous proof.

Let $\eta = 0.1, N = 2000$ and $w_0 = (1, 0)$. Figure 13 shows that the angles live in a small neighborhood of the assymptotically stable equilibrium point $(1, 0)$ on the circle. This is an illustration of the advantage of not working in $\mathbf{R}^2$, but on the circle S^1.

Figure 14 illustrates the same effect, when we use $\eta = 0.01, N = 2000$ and $w_0 = (1, 0)$. The weights again are living in a neighborhood of $(1, 0)$ and the corresponding separation lines in a neighborhood of the y-axis.

Figures 15 and 16 show the histograms of the angles, when $N = 5000$ and $w_0 = (1, 0)$ for respectivally $\eta = 0.1$ and $\eta = 0.01$. The angles are centered around 0.

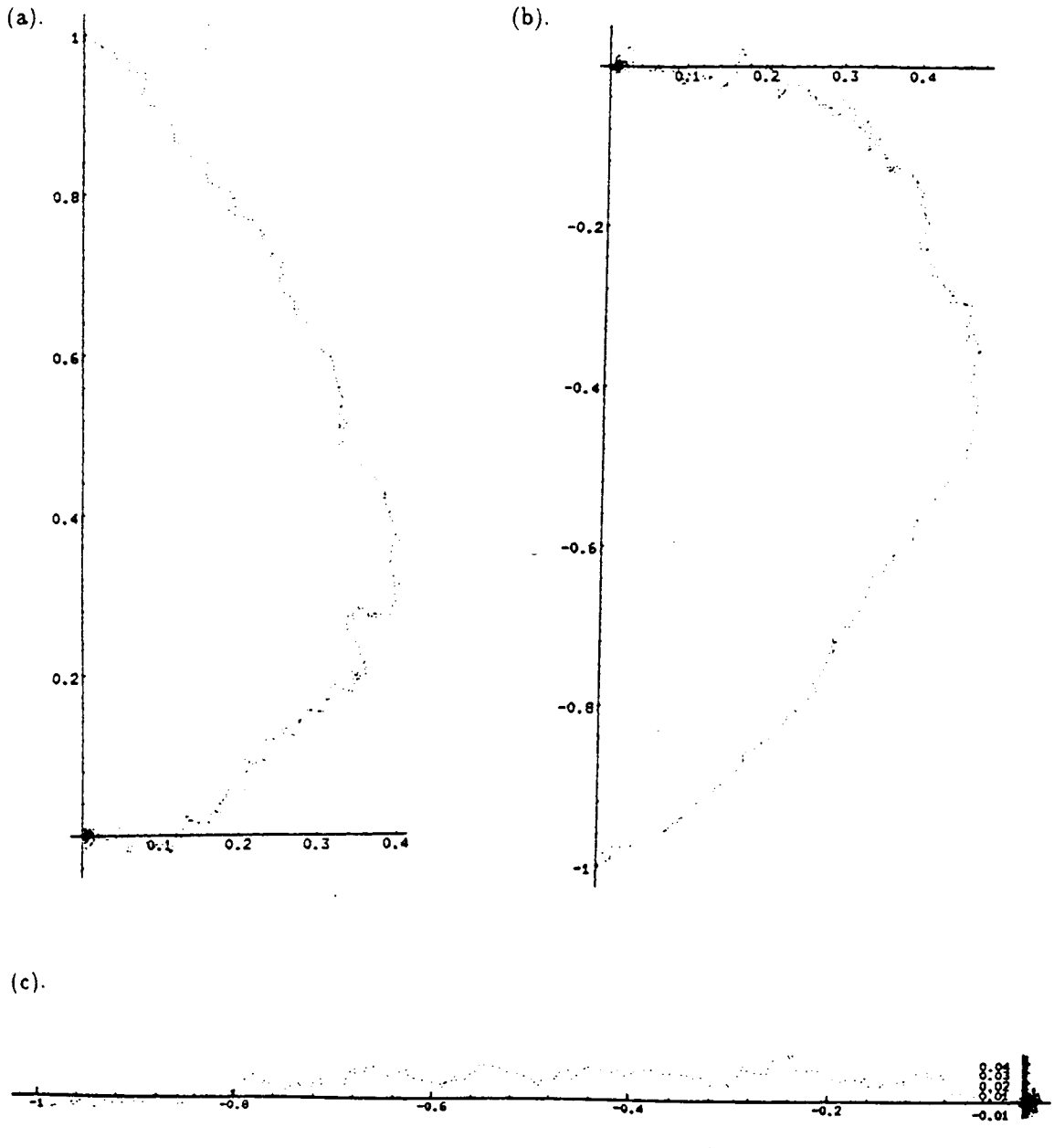

Figure 9. The weight process for $N = 2000, \eta = 0.01$ and (a).$w_0 = (0, 1)$, $(b).w_0 = (0, -1)$, $(c).w_0 = (-1, 0)$

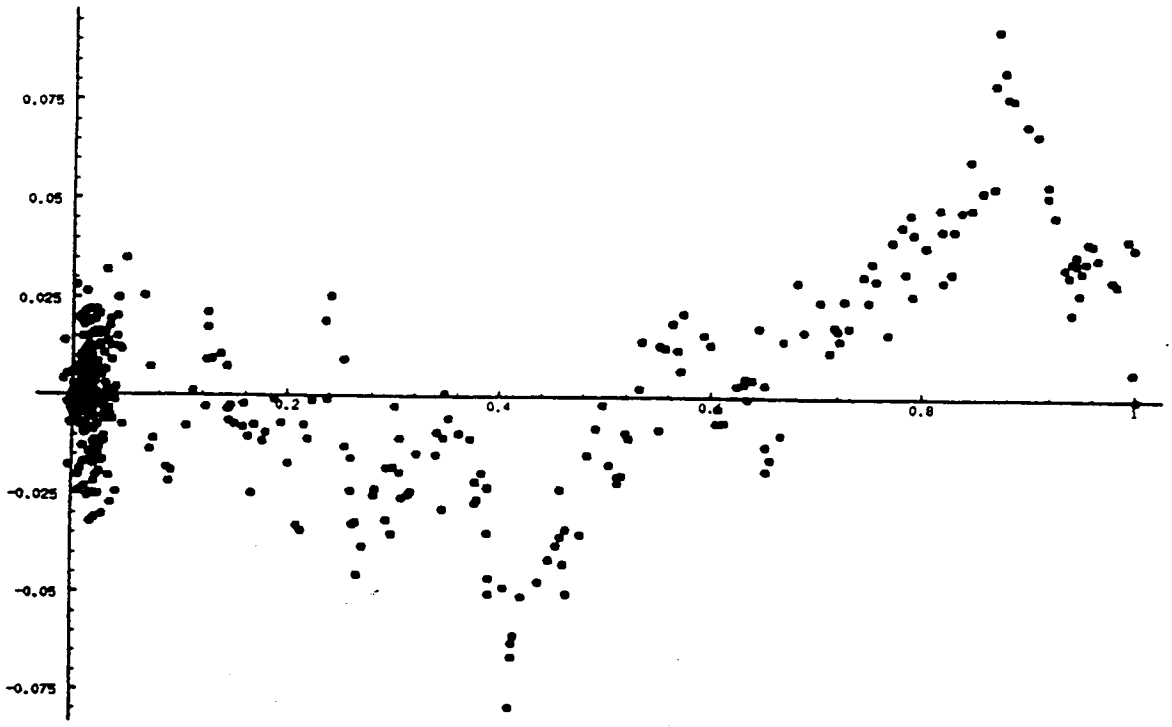

Figure 10. The weight process for $N = 2000$, $\eta = 0.01$ and $w_0 = (1, 0)$

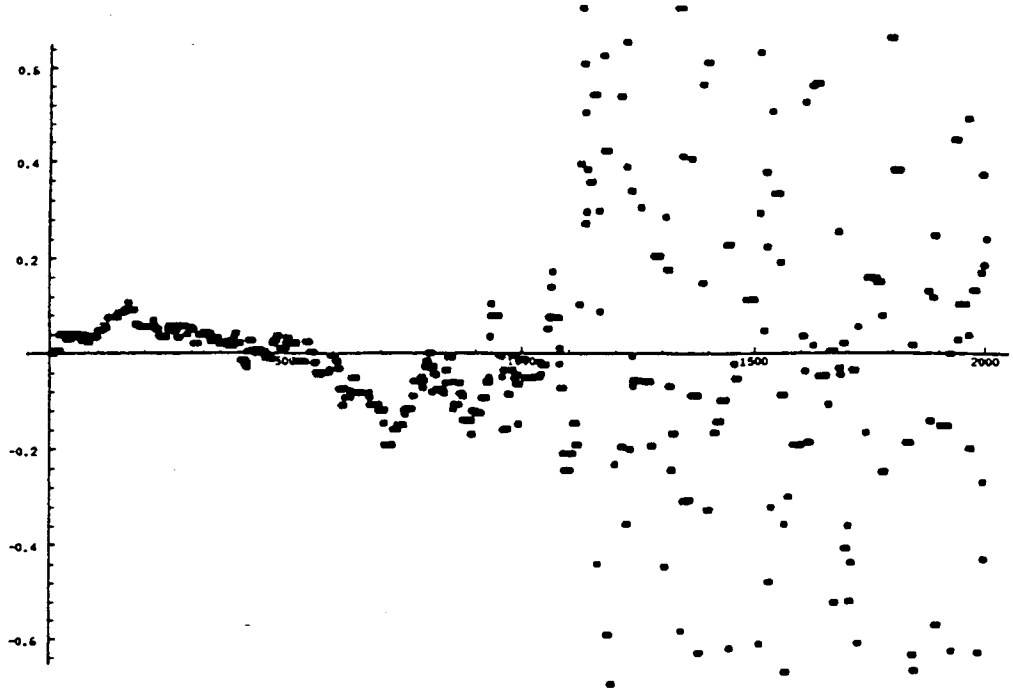

Figure 11. The graph of $\tan(\varphi_n)=\frac{w_n(2)}{w_n(1)}$ for $N = 2000$, $\eta = 0.01$ and $w_0 = (1, 0)$

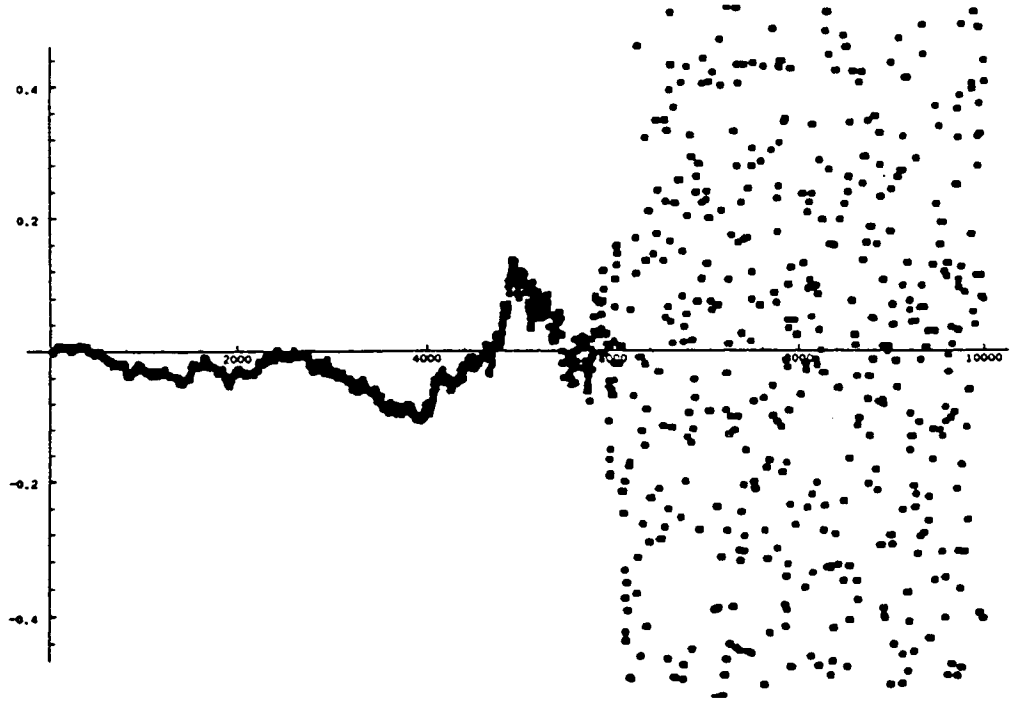

Figure 12. The graph of $\tan(\varphi_n)$ for $N = 10,000, \eta = 0.002$ and $w_0 = (1, 0)$

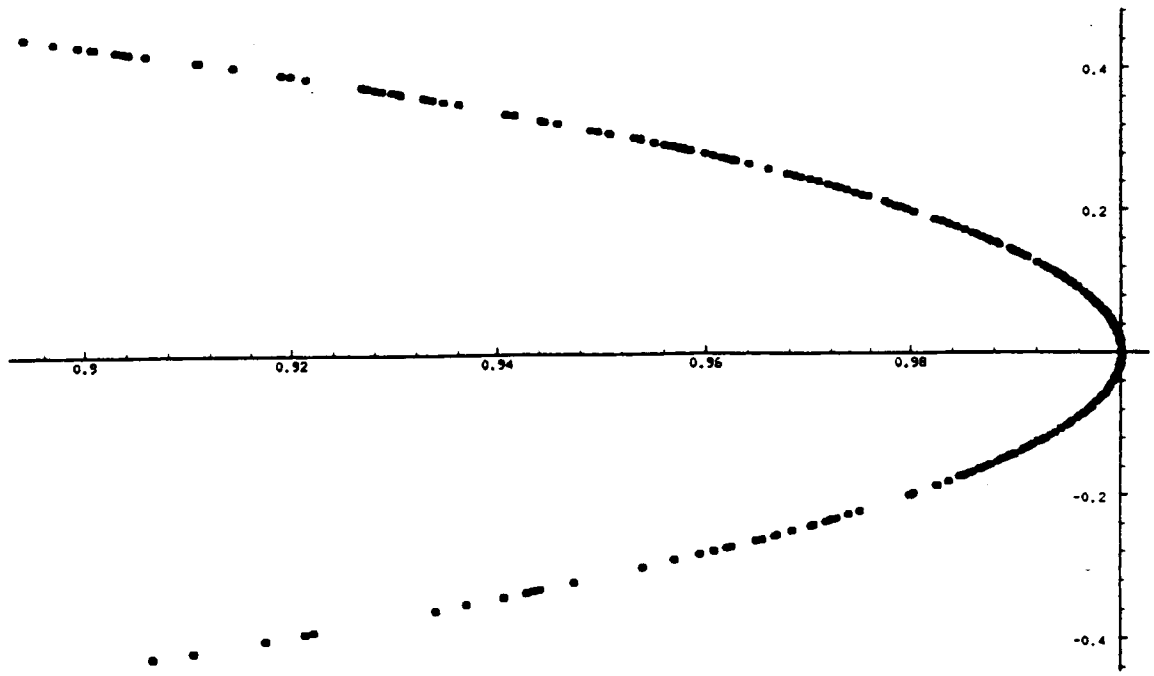

Figure 13. The weight process on S^1 for $N = 2000, \eta = 0.1$ and $w_0 = (1, 0)$

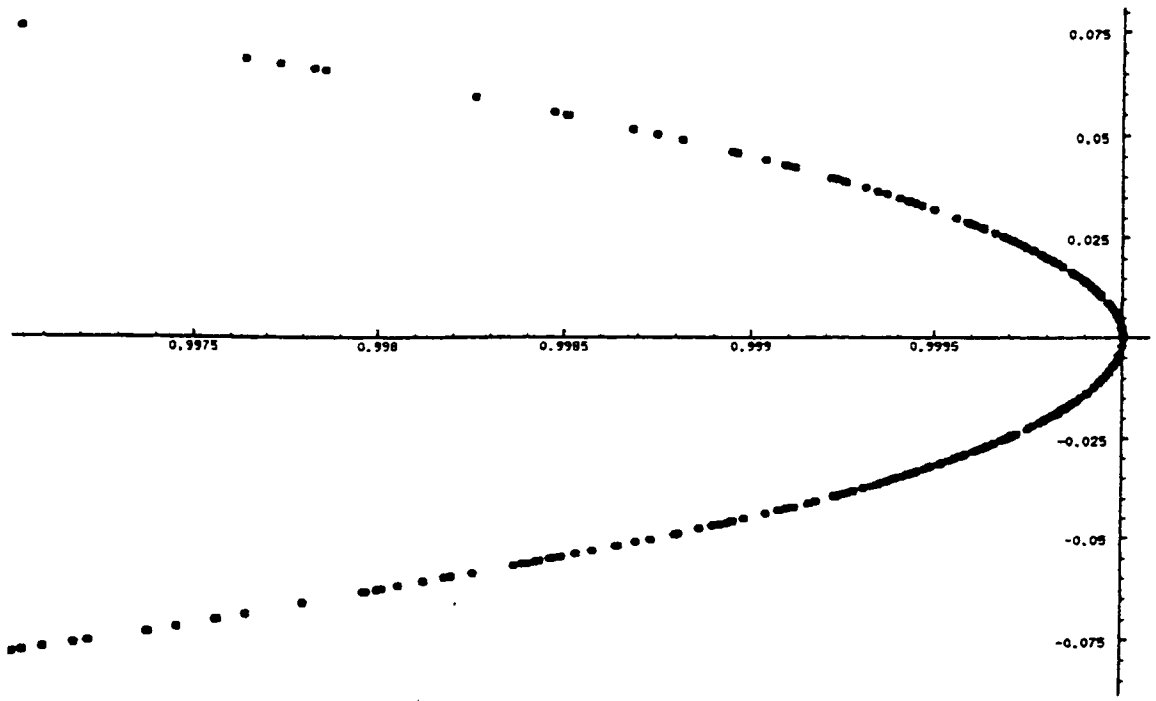

Figure 14. The weight process on S^1 for $N = 2000, \eta = 0.01$ and $w_0 = (1, 0)$

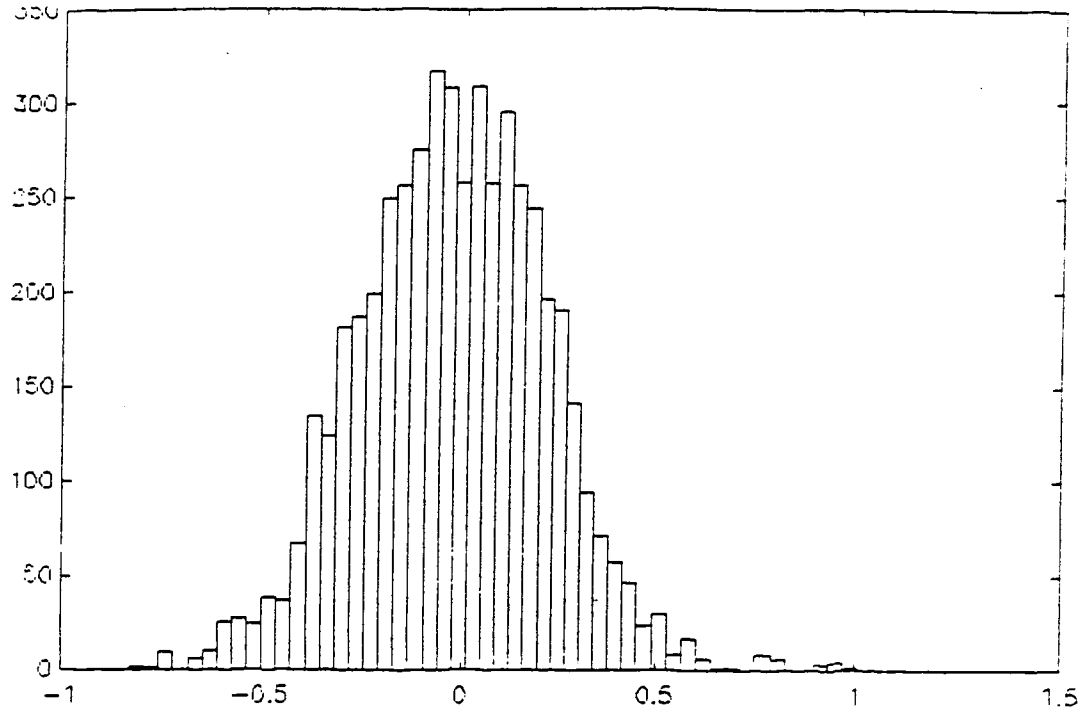

Figure 15. The histogram of the angles for $N = 5000, \eta = 0.1$ and $w_0 = (1, 0)$

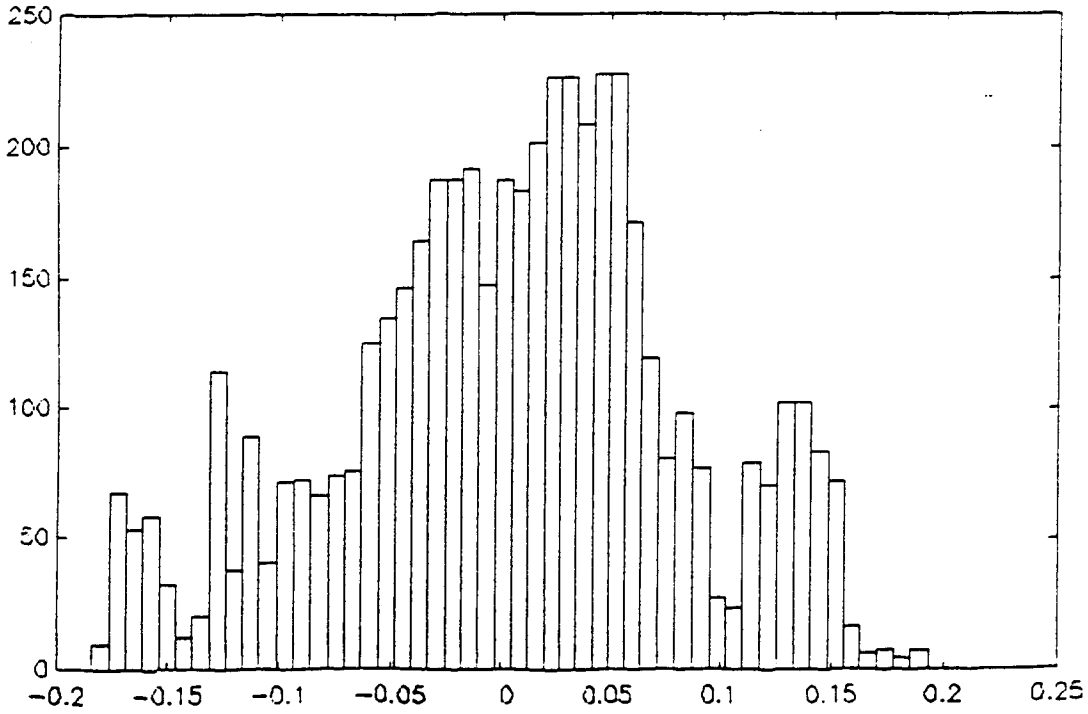

Figure 16. The histogram of the angles for $N = 5000, \eta = 0.01$ and $w_0 = (1, 0)$

Bibliography

[1]. R.M.Burton - Neural networks.
Lecture notes. T.U.Delft,The Netherlands. 1991

[2]. R.M.Burton,H.G.Dehling and U.Rösler - Small random perturbations of dynamical systems.
Preprint. University of Groningen,The Netherlands. 1993

[3]. J.L.Doob - Stochastic processes.
Wiley. New York. 1953

[4]. S.N.Ethier and T.G.Kurtz - Markov processes. Characterization and convergence.
Wiley. New York. 1986

[5]. T.Khanna - Foundations of neural networks.
Addison-Wesley. Reading,Massachusetts. 1990

[6]. W.S.McCulloch and W.Pitts - A logical calculus of the ideas immanent in nervous activity.
Bulletin of mathematical biophysics. Vol.5. 1943

[7]. M.Minsky and S.Papert - Perceptrons. An introduction to computational geometry.
The MIT Press. Cambridge, Massachusetts. 1969

[8]. Y. Pao - Adaptive pattern recognition and neural networks.
Addison-Wesley. Reading,Massachusetts. 1989

[9]. D.B.Pollard and R.L.Tweedie - R-theory for Markov chains on a topological state space.

J.London Math.Soc. 1975

[10]. F.Rosenblatt - Principles of neurodynamics.

Spartan Books. New York. 1962

[11]. R.Schalkoff - Pattern recognition. Statistical, structural and neural approaches.

Wiley. New York. 1992

[12]. R.L.Tweedie - R-theory for Markov chains on a general state space I: solidarity properties and R-recurrent chains.

Ann. Probab. 2. 1974

[13]. R.L.Tweedie - Sufficient conditions for ergodicity and recurrence of Markov chains on a general state space.

Stochastic Processes and their applications 3. 1975

[14]. P.D.Wasserman - Neural computing. Theory and practice.

van Nostrand Reinhold. New York. 1989